Questioning the Survey Responses of Large Language Models

Ricardo Dominguez-Olmedo* Moritz Hardt*‡ Celestine Mendler-Dünner*§‡

*Max Planck Institute for Intelligent Systems, Tübingen, Germany
§ELLIS Institute Tübingen, Germany
‡Tübingen AI Center

Abstract

As large language models increase in capability, researchers have started to conduct surveys of all kinds on these models in order to investigate the population represented by their responses. In this work, we critically examine language models' survey responses on the basis of the well-established American Community Survey by the U.S. Census Bureau and investigate whether they elicit a faithful representations of any human population. Using a de-facto standard multiple-choice prompting technique and evaluating 39 different language models using systematic experiments, we establish two dominant patterns: First, models' responses are governed by ordering and labeling biases, leading to variations across models that do not persist after adjusting for systematic biases. Second, models' responses do not contain the entropy variations and statistical signals typically found in human populations. As a result, a binary classifier can almost perfectly differentiate model-generated data from the responses of the U.S. census. At the same time, models' relative alignment with different demographic subgroups can be predicted from the subgroups' entropy, irrespective of the model's training data or training strategy. Taken together, our findings suggest caution in treating models' survey responses as equivalent to those of human populations.

1 Introduction

Surveys have a long tradition in social science research as a means for gathering statistical information about the characteristics, values, and opinions of human populations [Groves et al., 2009]. Insights from surveys inform policy interventions, business decisions, and science across various domains. Surveys typically consist of a series of well-curated questions in a multiple-choice format, with unambiguous framing and a set of answer choices carefully selected by domain experts. Surveys are then presented to groups of individuals and their answers are aggregated to gain statistical insights about the populations that these groups of individuals represent.

Many established survey questionnaires together with the carefully collected answer statistics are publicly available. Machine learning researchers have identified the potential benefits of building on this valuable data resource to study large language models (LLMs). Survey questions offer a way to systematically prompt LLMs, and the aggregate statistics over answers collected by surveying human populations serve as a reference point for evaluation. As a result, the use of surveys has recently gained popularity for studying LLMs' biases [Santurkar et al., 2023; Durmus et al., 2023]. At the same time, researchers in the social sciences have explored using LLMs to emulate the survey responses of specific human populations [Argyle et al., 2022; Lee et al., 2023]. If effective proxies, simulated responses could augment or replace the expensive data collection process of human responses.

It is tempting to prompt LLMs with survey questions, due to their syntactic similarity to question answering tasks [Brown et al., 2020; Liang et al., 2022]. However, it is a priori unclear how to

interpret their answers. Rather than knowledge testing, surveys seek to elicit aggregate statistics over individuals, providing an unbiased view on the properties of the population they represent. The quality of survey data hinges on the validity and robustness of the conclusions that can be drawn from it. Clearly, running a survey on LLMs is different from interrogating humans and comes with distinct challenges. While much research has gone into carefully designing surveys to ensure faithful human responses, it is unclear whether prompting LLMs with the same surveys satisfies similar premises out-of-the-box. We devote this work to gain insights into LLMs' survey responses, what we can expect to learn from them, and to what extent they resemble those of human populations.

1.1 Our work

The basis of our investigation is the American Community Survey[1] (ACS), a demographic survey conducted by the U.S. Census Bureau at a national level, on a yearly basis. We curate a questionnaire by selecting 25 multiple choice questions from the 2019 ACS. We prompt 39 language models of varying size with these questions, individually and in sequence, and we record their probability distribution over answers. Based on the collected data, we investigate:

1. What can we learn about LLMs, and their relative differences, from their survey responses?
2. Does the data generated by prompting models to answer the ACS questionnaire resemble the census data collected by surveying the U.S. population?

We start by inspecting models' distributions over answers to individual survey questions when the questions are asked independently. We observe that the entropy of models' responses differs substantially across models of varying size. We find that this differences arise because strong ordering and labeling biases confound models' answers. In fact, after adjusting for such systematic biases through randomized choice ordering, we find that response distributions are very similar across models and tend to correspond to highly balanced answers.

Comparing models' responses to those of the U.S. census population, we find that natural variations in entropy across questions are not reflected in models' responses. Instead, on average across questions, models' responses are no closer to the census population, or the population of any state within the US, than to a fixed uniform baseline. As a result, the relative alignment of model responses with different demographic subgroups can be explained by the entropy of the subgroups' responses, irrespective of the data or training procedure employed to train the model. We verify that our findings generalize beyond the ACS to other surveys considered by prior work, thus providing important context to prior studies that employ surveys to examine the biases of LLMs.

Lastly, we prompt LLMs to repeatedly answer entire survey questionnaires, generating for each model a synthetic tabular dataset which emulates in form the ACS census data. We present questions in a sequential manner, keeping a model's previous answers in context when prompting it to answer subsequent questions of the questionnaire. We then investigate whether these synthetic datasets resemble the responses collected by the U.S. Census. We use the *discriminator test* as an investigative tool: we construct a binary prediction task aiming to discriminate between model-generated data and the U.S. census data. We show that this prediction task can be solved with very high accuracy ($\geqslant 97\%$) for all models up to 70B parameters. For the latest GPT-4 model the accuracy is 92%, indicating that there is a substantial gap between model-generated data and human responses.

[1] https://www.census.gov/programs-surveys/acs

More broadly, our findings suggest caution when treating language models' survey responses as a faithful representation of any human population, putting the burden of proof on researchers wishing to draw valid conclusions about human populations from data collected by surveying LLMs.

1.2 Related work

Despite the syntactical similarities, evaluating LLMs on the basis of their survey responses differs from traditional question answering evaluations [Liang et al., 2022]. Question answering (QA) tasks predominantly serve the purpose of knowledge testing [e.g., Kwiatkowski et al., 2019; Rajpurkar et al., 2016; Talmor et al., 2019; Mihaylov et al., 2018]. In such setting, a language model's answer to some unambiguous input question is extracted by computing its most likely completion. Alternatively, models' most likely response to questions that lack a clear answer (e.g., "Angela and Patrick are sitting together. Who is an entrepreneur?") have been used to investigate various biases of LLMs [Li et al., 2020; Mao et al., 2021; Perez et al., 2022; Abid et al., 2021; Jiang et al., 2022].

When evaluating LLMs on the basis of survey questions, it is not models' most likely completion that is studied, but rather models' probability distribution over various answer choices. Santurkar et al. [2023] study LLMs' answer distributions for multiple-choice opinion polling questions, measuring their similarity to those of various U.S. demographic groups. They extract models' answer distributions from the next token probabilities corresponding to each answer choice. Durmus et al. [2023] employ a similar methodology but instead consider transnational opinion surveys. We adopt this popular methodology to investigate the properties of models' answer distributions on the basis of a well-established demographic survey, beyond measuring the relative similarity of models' responses to the survey responses of different human populations.

In addition to asking questions individually, we also prompt models to complete entire survey questionnaires. We present questions in a sequential manner, keeping a model's previous answers in context when prompting the model to answer subsequent questions. This methodology resembles prior work by Hartmann et al. [2023]; Rutinowski et al. [2023]; Motoki et al. [2023]; Feng et al. [2023] who sequentially prompt language models to answer entire political compass or voting advice questionnaires. But instead of aggregating answers into a political affinity score, our focus is on examining whether models' responses resemble those of human populations.

Lastly, there is an emerging body of research that integrates LLMs into computational social science [Ziems et al., 2023]. This includes tasks such as taxonomic labeling, where language models are employed for tasks such as opinion prediction [Kim and Lee, 2023; Mellon et al., 2022], and free-form coding, where language models are used to generate explanations for social science constructs [Nelson et al., 2021]. Recent studies have also investigated the feasibility of using LLMs to simulate human participants in psychological, psycholinguistic, and social psychology experiments [Dillion et al., 2023; Aher et al., 2023], or as proxies for specific human populations in social science research [Argyle et al., 2022; Lee et al., 2023; Sanders et al., 2023] and economics [Brand et al., 2023; Horton, 2023]. Within this context, our work suggests caution in relying on the survey responses of LLMs to elicit synthetic responses that resemble those of human populations.

2 Surveying language models

We employ the de-facto standard methodology to survey language models introduced by Santurkar et al. [2023]. For every survey question, we collect language models' probability distribution over answer choices. Formally, for a given model m and survey question q we define the model's *survey*

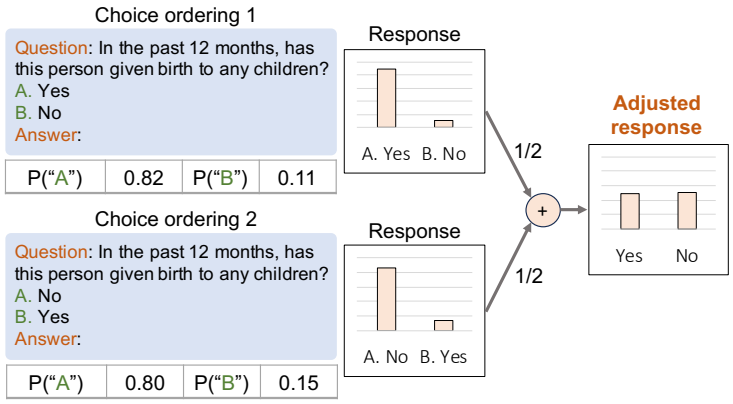


Figure 1: For any given question, a model's response distribution is computed using its next token probabilities for each choice label. When adjusting for choice ordering bias, we average the model's responses over all possible answer choice orderings.

response as a categorical random variable R_q^m which can take on k_q values corresponding to the number of answer choices to question q. We determine the event probabilities of R_q^m by prompting model m as follows:

1. We construct an input prompt of the form "`Question: <question> \n A. <choice 1 >\n B. <choice 2> \n ... <choice` k_q`> \n Answer:`".
2. We query language models with the input prompt and obtain their output distribution over next-token probabilities. We select the k_q output probabilities corresponding to each answer choice (e.g., the tokens "A", "B", etc.), and we renormalize to obtain the probability distribution over survey answers. For OpenAI's models, we only have access to the top-5 next-token log probabilities through the OpenAI API. In this case, we assign to the unseen probabilities (if any) the minimum between the remaining probability mass and the smallest observed probability, following the methodology of Santurkar et al. [2023].

The chosen style of prompt is standard for question answering tasks [Hendrycks et al., 2021], used in OpinionQA [Santurkar et al., 2023], and follows the best practices for social science research recommended by Ziems et al. [2023]. For completeness we perform several prompt ablations, including the prompt variations used by Argyle et al. [2022], Santurkar et al. [2023] and Durmus et al. [2023]. We find our take-aways to be robust to such changes, see Appendix D.

Survey questions. We use a representative subset of 25 multiple-choice questions from the 2019 ACS questionnaire. We denote the set of questions by Q. The questions cover basic demographic information, education attainment, healthcare coverage, disability status, family status, veteran status, employment status, and income. We generally consider the questions and answers as they appear in the ACS questionnaire. We refer to Appendix A.1 for the exact framing we used for each question.

Models surveyed. We survey 39 language models of size varying from 110M to 175B parameters: the base models GPT-2 [Radford et al., 2019], GPT-Neo [Black et al., 2021], Pythia [Biderman et al., 2023], MPT [MosaicML, 2023], LLaMA [Touvron et al., 2023a], Llama 2 [Touvron et al., 2023b], and GPT-3 [Brown et al., 2020]; as well as the instruct variants of MPT 7B and GPT NeoX 20B,

the Dolly fine-tune of Pythia 12B [Databricks, 2023], the Vicuna and Koala fine-tunes of LLaMA 7B and 13B [Geng et al., 2023; Chiang et al., 2023], Llama 2 Chat [Touvron et al., 2023b], GPT-3.5, GPT-4 [OpenAI, 2023], and the text-davinci variants of GPT-3 [Ouyang et al., 2022].

Reference data & evaluation. We use the responses collected by the U.S. Census Bureau when surveying the U.S. population as our reference data. In particular, we use the 2019 ACS public use microdata sample[2] (henceforth census data). The data contains the anonymized responses of around 3.2 million individuals in the United States. For each survey question $q \in Q$, we denote the census' population-level response as a categorical random variable C_q whose event probabilities are the relative frequency of each answer choice among survey respondents. We use U_q to denote the uniform distribution over answers. Given these two reference points, we evaluate language models' responses R_q^m along two dimensions:

- We use *entropy* to measure the degree of variation in models' responses. We denote the entropy of a random variable R as $H(R)$. To meaningfully compare the entropy of responses across questions with varying number of choices k_q, we report normalized entropy, that is, the entropy relative to the uniform distribution. $H(R_q^m) = 1$ implies that model m's survey response to question q is uniformly distributed (i.e., $H(U_q) = 1$).
- We use the *Kullback–Leibler (KL) divergence* to measure the "similarity" between two distributions over answers. We write $\mathrm{KL}(R_q^m \,||\, C_q)$ for the KL divergence between the response distribution R_q^m of model m to question q and the corresponding aggregate response distribution C_q observed in the census data. The larger the KL distance between two distributions, the more dissimilar the two distributions are.

Randomized choice ordering. For several investigations we survey models under randomized choice ordering. For a given question q, we prompt models with different permutations of the answer choice ordering, i.e., the assignment of answers (e.g., "male", "female") to choice labels ("A", "B", etc), while the order of choice labels is kept alphabetically. We evaluate models' survey responses under all possible choice orderings and we use $\bar{R}_q^m$ to denote the expected distribution over answers and $\bar{O}_q^m$ to denote the expected distribution over selected choice labels. For questions with more than 6 answers we evaluate a maximum of 5000 permutations. For OpenAI's models we evaluate up to 50 permutations due to the costs of querying the OpenAI API. This distinction serves to decouple a model's tendency towards picking a particular answer from its tendency towards picking a particular choice label. In the following we refer to the expected survey response $\bar{R}_q^m$ under uniformly distributed choice ordering as the *adjusted* survey response. See Figure 1 for an illustration of the methodology used for adjusting for choice ordering response biases.

3 Inspecting models' survey responses

We start by surveying the base pre-trained models. We present survey questions independently of one another, showing the answer choices in the same order as the ACS.

For a first investigation, we consider the normalized entropy of models' responses to the "SEX", "HICOV", and "FER" questions. The SEX question inquiries about the person's sex, encoded as male female, the HICOV question inquiries whether the person is currently covered by any health insurance

[2] https://www.census.gov/programs-surveys/acs/microdata

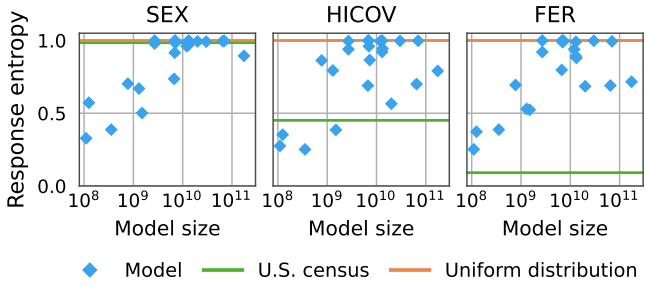
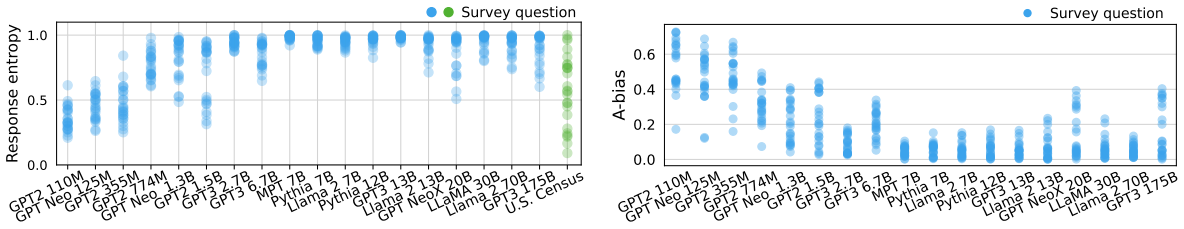


Figure 2: Entropy of responses across model sizes for the SEX, HICOV, and FER survey questions. Entropy of models' responses (◆) tends to increase log-linearly with model size, irrespective of the underlying response entropy observed in the U.S. census (–).



(a) Entropy of base models' responses.

(b) A-bias of base models' responses.

Figure 3: Entropy and A-bias of models' responses across ACS questions, where each dot corresponds to one of the 25 questions. Models are ordered by size. (a) Models' variation in response entropy across questions is much smaller than the variation observed in the census data, shown in green. (b) All models suffer from substantial A-bias, particularly the smaller models.

plan, and the FER question inquires whether the person has given birth in the past 12 months. When surveying the U.S. population, these three questions elicit responses with very different entropy; responses to the SEX question are almost uniformly distributed, whereas most people answer "No" to the FER question. In contrast, as shown in Figure 2, the entropy of models' responses to these three questions are surprisingly similar. In particular, we find that the entropy of models' responses tends to increase log-linearly with model size. This trend is consistent across all ACS survey questions, see Figure 9 in Appendix B.1.

For a broader picture, we illustrate models' response entropy across all survey questions in Figure 3(a). The blue dots represent models' responses to individual questions, and the green dots represent the entropy of the responses of the U.S. census. We order models by size. We observe that the entropy of responses of the U.S. census greatly varies across questions. In contrast, for any given model, the entropy of its responses varies substantially less so.

Overall, we find that models' response distributions seem to be widely independent of the survey question asked, and variations across models are much larger than variations across questions. This lead us to suspect that variations across models might arise mostly due to systematic biases.

3.1 Testing for systematic biases: A-bias

It is well-known that language models' most likely answer to multiple-choice questions can change depending on seemingly minor factors such as the ordering of few-shot examples [Zhao et al., 2021; Lu et al., 2022] or the ordering of answer choices [Robinson and Wingate, 2023]. We are interested

in the extent to which changes in choice ordering affect a model's output *distribution over answers*.

We start by measuring *A-bias*: the tendency of a model towards picking the answer choice labeled "A". In particular, we seek to study the extent to which the strength of this bias explains the differences in responses observed across models. For an unbiased model that outputs the same answer distribution irrespective of choice ordering, the expected choice distribution $\bar{O}_q^m$ under randomized choice ordering would match precisely the uniform distribution (e.g., P("A") = P("B") = 0.5). We define a model's A-bias as its absolute deviation from this unbiased baseline:

$$\text{Abias}_q^m := \left| \text{P}(\bar{O}_q^m = \text{“}A\text{”}) - 1/k_q \right| \tag{1}$$

We measure A-bias for each question q and model m. Results are illustrated in Figure 3(b). We sort models by their size. We observe all models exhibit substantial A-bias. However, models in the order of a few billion parameters or fewer consistently exhibit particularly strong A-bias, and tend towards mono answers. We additionally observe that the strength of A-bias in instruction or RLHF tuned models is similar to that of base models, see Appendix B.2.

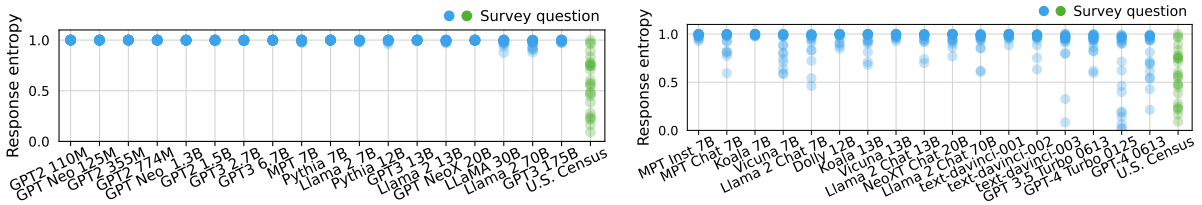
We investigate other types of labelling and position bias (e.g., last-choice bias) in Appendix C. Overall, we find a strong tendency of LLMs to pick up on spurious signals in the way that answers are ordered and labeled, rather than their semantic meaning. Notably, in contrast to the primacy bias observed in humans [Groves et al., 2009], we find that models exhibit substantial A-bias even when randomizing the position of the "A" choice. Our findings are consistent with the concurrent work of Tjuatja et al. [2023], which similarly finds that models' response biases to multiple-choice survey questions are generally not human-like.

In summary, we find that systematic biases confound models' answer distributions. This makes it challenging to draw robust conclusions about general properties of LLMs, and their comparison, from survey responses. For example, simply reversing the order of answers to the "SEX" question could lead to GPT-2 seemingly representing a population where females are significantly over-represented, whereas a reverse conclusion would be drawn when using the standard answer order. While much research went into designing the ACS to elicit faithful answers and eliminate systematic biases when surveying human populations, simply using the same question framing does not protect against the systematic response biases that language models exhibit.

4 Controlling for labeling and ordering biases

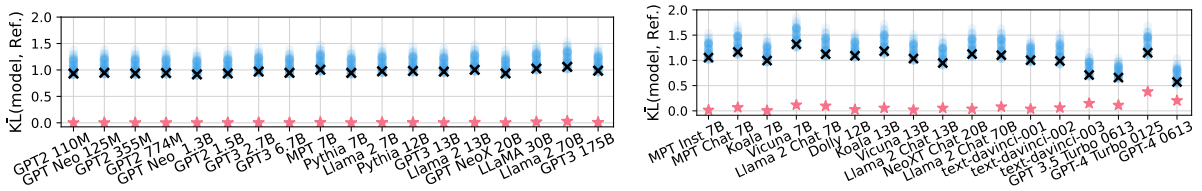
To eliminate confounding due to labeling and ordering biases, we survey models under randomized choice ordering, borrowing an established methodology to adjust for ordering biases of all kinds in survey research [Groves et al., 2009]. In the following, we refer to the expected response after answer choice randomization as the *adjusted* response.

In Figure 4(a) we plot the normalized entropy of base models' adjusted responses for the ACS questions considered. We find that after adjustment, 1) the variations in responses' entropy across survey questions are very small, 2) we no longer observe the trend of the entropy of model responses increasing log-linearly with model size. In fact, models' survey responses have a normalized entropy of approximately 1 irrespective of model size or survey question asked. This validates our initial hypothesis that, without adjustment, variations in responses across base models arise predominantly due to systematic biases such as A-bias, rather than the content of the survey questions asked.



(a) Base models. (b) Instruction-tuned models.

Figure 4: Entropy of models' responses after adjusting for choice ordering biases. Dots correspond to individual questions and models are ordered by size. (a) The entropy of base models' adjusted responses is close to 1 (i.e., uniform). (b) Instruction tuned-models exhibit substantially higher variations in entropy across questions.



(a) Base models (b) Instruction-tuned models

Figure 5: Divergence between adjusted model responses and different baselines: the overall U.S. census (✖), individual U.S. states (●), and a uniform baseline with uniformly random responses (★). Model responses are consistently more similar to the uniform baseline than to any human reference population.

4.1 Effect of instruction tuning

We now evaluate language models that have been fine-tuned with instructions and/or human preferences, henceforth "instruction-tuned models". In Figure 4(b) we plot the normalized entropy of instruction-tuned models' ACS survey responses after adjustment. We observe that instruction tuned-models all exhibit substantially higher variations in entropy across questions compared to base models. Nonetheless, the entropy of their responses is nonetheless higher than the entropy of the human responses observed in the U.S. census data.

4.2 Comparing model responses to the U.S. census

We now investigate the similarity of language models' adjusted responses to the census data. To do so, we consider the overall U.S. census population, as well as 50 census subgroups corresponding to every state in the United States.

Inspired by the alignment measures proposed by Santurkar et al. [2023] and Durmus et al. [2023], we investigate the similarity of model responses to the census data by evaluating the average divergence across questions between model responses and the census statistics.[3] We evaluate average KL divergence between each language model m and each reference population Ref, as follows:

$$\bar{\mathrm{KL}}(m, \mathrm{Ref}) = \frac{1}{|Q|} \sum_{q \in Q} \mathrm{KL}(\bar{R}_q^m || \mathrm{Ref}_q).$$

[3]Whereas Santurkar et al. [2023] use the Wasserstein distance to compare answer distributions, we use KL divergence since questions in the ACS are predominantly nominal, rather than ordinal.

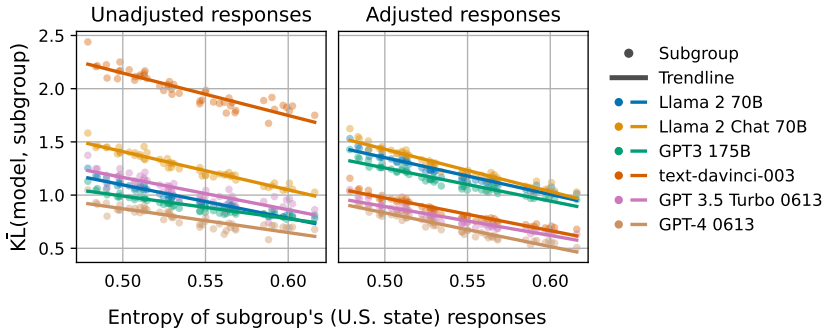


Figure 6: Alignment of models with different census subgroups. All models tend to exhibit similar relative alignment, and this alignment is correlated with the entropy of the subgroups' responses.

Results are depicted in Figure 5. For each model we plot the divergence to the census in black, the divergence to the different subgroups in blue, and the divergence to a uniform baseline with balanced responses in red. We observe that models are strikingly more similar to the uniform baseline than to any of the populations considered. For base models, this result is unsurprising, since in the previous section we established that base models' responses are essentially uniform after adjustment.

Looking at Figure 5(b) we find no consistent trend that instruction-tuning would move responses closer to the census, despite the increased deviation from uniform and the larger variations in entropy (recall Figure 4(b)). Only for larger models the divergence seems to clearly decrease with instruction-tuning. However, all models' responses still remain significantly closer to the uniform baseline than to the U.S. census. For instance, the model with the largest number of responses being more similar to those of the U.S. census than to uniform is GPT-4, for which 6 out of 25 questions (24%) are closer to the U.S. census than to the uniform baseline.

4.3 Implications for survey-based alignment metrics

Our findings add important context to previous works studying the relative similarity between the survey responses of models and those of specific populations. In particular, due to the models' tendency towards balanced answers, there is a strong correlation between a models' similarity (or alignment) with a subgroup and the subgroup's entropy, as shown in Figure 6. While we only depict the largest models in this figure, this correlation consistently holds for all models surveyed (see Appendix B.3). Interestingly, this trend also holds pre-adjustment, and we find that only the relative ordering of models changes with adjustment.

Thus, our survey-derived alignment measure is more informative of the reference populations rather than the language models they aim to evaluate, since variations across subgroups tend to be more pronounced than variations across models. Particularities, such as the training data used or the demographics of the annotators used for fine-tuning with human feedback, seems to have little impact on which population is best represented. Models therefore consistently appear to be more "aligned" with the subpopulations exhibiting high entropy in their answers.

Beyond the ACS. To inspect whether this trend changes with the content of the questions asked, we reproduce our experiments with additional surveys. We use the American Trends Panel (ATP) opinion surveys considered by Santurkar et al. [2023], and the Pew Research's Global Attitudes Surveys (GAS) and World Values Surveys (WVS) considered by Durmus et al. [2023]. These surveys encompass

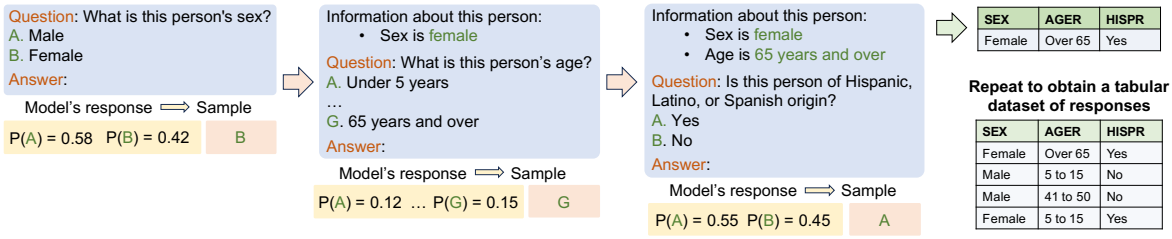


Figure 7: Methodology and prompt template used to sequentially sample models' responses to entire survey questionnaires. We provide the answers to previous question in context when prompting subsequent questions. The output is a tabular dataset of responses.

around 1500 questions and 60 U.S. demographic subgroups, and around 2300 questions and 60 national populations, respectively. We adopt the alignment metrics considered by the aforementioned works. We find that the insights gained from the ACS also hold for the ATP and GAS/WVS surveys.

In particular, we similarly find a linear trend between the alignment metrics and subgroups' entropy of responses, see Figure 18 in Appendix E.3. This observation explains some of the findings in prior works. For example, Santurkar et al. [2023] find that "all the base models share striking similarities–e.g., being most aligned with lower income, moderate, and Protestant or Roman Catholic groups" and "our analysis [...] surfaces groups whose opinions are poorly reflected by current LLMs (e.g., 65+ and widowed individuals)". For the ATP surveys considered, low income, moderate, and Protestant/Catholic are precisely the demographic subgroups with responses closest to uniformly random among the income, political ideology, and religion demographic subgroups; whereas age 65+ and widowed are the demographic subgroups with responses furthest from uniform among the age and marital status demographic subgroups.

5 Responses to entire questionnaires

In the previous sections, survey questions were presented independently of one another. We now seek to fill entire ACS questionnaires in a sequential manner, in order to generate for each language model a synthetic dataset of responses which emulates in form the ACS dataset collected by the U.S. Census Bureau. We then study the extent to which such synthetic datasets resemble the ACS dataset.

5.1 Methodology

We present survey questions in the same order as in the ACS questionnaire. When querying a model to answer survey question q, we include a summary of the $q-1$ previously sampled answers in context. [4] We then sample from the model's output probability distribution over answers, and continue to the next question. We illustrate this sequential process in Figure 7. We refer to Appendix D.3 for results collected with different variations of how a model's previous answers are integrated into the prompt. We find our results to be robust to these prompt variations.

For each language model we sample N=100,000 model-generated responses to the ACS. Due to the cost of querying OpenAI's models, we only survey GPT-4 and sample $N = 500$ responses. As a result, we generate for each language model a tabular dataset similar in form to the ACS data, with N rows corresponding to each filled questionnaire and 25 columns corresponding to each question.

[4] The maximum number of tokens in a filled questionnaire was less than 1024 tokens in all cases, thus fitting entirely within the context window of all surveyed models.

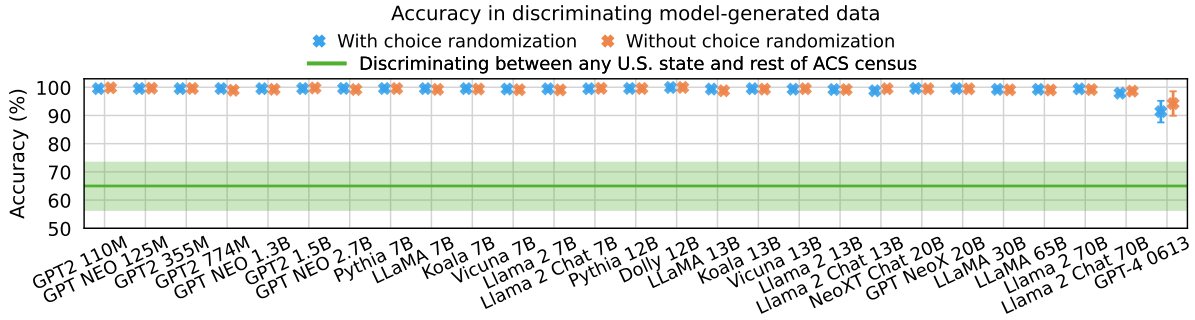


Figure 8: Accuracy of the discriminator test. For all language models, it is possible to discriminate with very high accuracy between the ACS census data and model-generated data, (✖) before adjustment and (✖) after adjustment. We contrast this against the accuracy value of discriminating between the ACS data of any given U.S. state and the rest of the ACS census data (–).

5.2 The discriminator test

We investigate whether the model-generated datasets resemble the U.S. census data by constructing a binary prediction task aiming to discriminate synthetic responses from census responses. Intuitively, if the two datasets were very dissimilar, then a classifier would be able to achieve high accuracy. Formally, let $\mathscr{F}$ be class of binary prediction functions mapping each data point (i.e., a row in the tabular dataset) to $\{0, 1\}$, then the accuracy of the best $f \in \mathscr{F}$ on the discriminator task provides a lower bound on the total variation (TV) distance between the two empirical data distributions.

Hence, we train a predictor f to discriminate between the model-generated data and the census data in order to obtain an empirical lower bound on the distance between the two datasets. Specifically, we concatenate to each model-generated dataset a random sample of N individuals from the ACS census data, and introduce a binary label indicating whether each row of the concatenated dataset was model-generated or not. We then train an XGBoost classifier in this binary prediction task. As an additional point of reference, we also consider the accuracy in discriminating between the census data of any given U.S. state and an equally-sized sample of the ACS data of all other U.S. states.

We report mean test accuracy in Figure 8. We consider 100 different random seeds. We find that the trained classifiers can differentiate between model-generated data and census data with very high accuracy ($> 90\%$) in all cases. Therefore, the empirical distributions corresponding to the model-generated data and the census data have TV distance larger than 0.9. These stark results indicate that data generated by sequentially prompting language models with the ACS survey questionnaire bears little similarity with the data collected by surveying the U.S. population.

5.3 Contrast with silicon samples

Argyle et al. [2022] propose "silicon sampling", a methodology to produce synthetic survey respondents using LLMs by conditioning on actual survey respondents. They focus on a subset of 12 questions from the 2016 American National Election Studies (ANES) survey. For every human respondent, they construct a corresponding "silicon individual" by querying GPT-3 to predict the ANES respondent's answer to each survey question given the respondent's answers to all other questions. Their results indicate that, for the 2016 ANES survey, GPT-3 can be a fairly calibrated predictor of an individual's

answer to some survey question conditioned on the respondent's answers to all other survey questions.[5]

However, Argyle et al. [2022] emphasize that important insights can be gained by emulating the survey responses of human populations "prior to or in the absence of human data". In this work we have considered precisely the setting where models' responses are obtained in the absence of human data.[6] To investigate how our findings transfer to the ANES, we reproduce the experiments of Section 5 using the 2016 ANES survey questionnaire considered by Argyle et al. [2022] and their "interview-style" prompt. We apply the discriminator test, and find that the trained classifiers can discriminate between the model-generated data and the ANES data with accuracy > 99 % (see Appendix E), indicating that models' responses are markedly different to those in the ANES data.

Thus, the fact that models may perform reasonably well at feature imputation tasks (e.g., *predicting* an individual's answer to some question given their answers to *all other questions*) does not imply that models can *generate* synthetic respondents that resemble the responses obtained by surveying human populations. This suggests caution when using LLMs to emulate human populations at present time, in particular in the absence of human data.

6 Conclusion

We closely examined the survey responses of LLMs on the basis of the prime demographic survey conducted in the United States. To do so, we leveraged a popular methodology to elicit LLMs' answer distributions to survey questions. We found that model responses are dominated by systematic ordering biases and do not exhibit the natural variations in entropy found in the human reference data collected by the US census. Even after adjusting for ordering biases, LLMs' responses still do not resemble those of human populations. Instead, they exhibit consistently high entropy. This holds true both when presenting questions individually and in sequence, and irrespective of model size or fine-tuning with human preferences.

Taken together, our findings caution to expect robust insights when comparing LLMs' responses against those of humans. In our study we could not find any indication that LLMs elicit faithful representations of any human population. It remains to be seen how to take advantage of the flexibility, scalability and scope of LLM survey data in the context of social science research. What seems clear is that LLMs' survey data should be treated differently to human data and that the validity of surveys as a measurement instrument should be readdressed. The robustness and quality of an established survey does not seamlessly translate from the results obtained by surveying human populations to the logits output by LLMs.

[5] Lee et al. [2023] and Sanders et al. [2023] study imputation tasks similar to those of Argyle et al. [2022], and find that LLMs are not calibrated predictors for a variety of such tasks.

[6] Conceptually, to generate each synthetic individual, we start with a blank survey questionnaire and prompt the LLM to sequentially fill the entire questionnaire. Whereas we only prompt LLMs with their own responses (i.e., to previous survey questions), Argyle et al. [2022] prompt the model only with actual human responses from the 2016 ANES data.

References

Abid, A., Farooqi, M., and Zou, J. (2021). Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Aher, G. V., Arriaga, R. I., and Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C., and Wingate, D. (2022). Out of one, many: Using language models to simulate human samples. *arXiv preprint arXiv:2209.06899*.

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and van der Wal, O. (2023). Pythia: A suite for analyzing large language models across training and scaling. *arxiv prepring arxiv:2304.01373*.

Black, S., Gao, L., Wang, P., Leahy, C., and Biderman, S. (2021). GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Brand, J., Israeli, A., and Ngwe, D. (2023). Using GPT for Market Research. *Harvard Business School Marketing Unit Working Paper No. 23-062*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Databricks (2023). Dolly 12b.

Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*.

Dorner, F., Sühr, T., Samadi, S., and Kelava, A. (2023). Do personality tests generalize to large language models? In *NeurIPS Workshop on Socially Responsible Language Modelling Research*.

Durmus, E., Nyugen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., et al. (2023). Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Feng, S., Park, C. Y., Liu, Y., and Tsvetkov, Y. (2023). From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. *Findings of the Association for Computational Linguistics: ACL 2023*.

Geng, X., Gudibande, A., Liu, H., Wallace, E., Abbeel, P., Levine, S., and Song, D. (2023). Koala: A dialogue model for academic research. Blog post.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., and Tourangeau, R. (2009). *Survey Methodology*. Wiley.

Hartmann, J., Schwenzow, J., and Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *NBER Working Paper*.

Jiang, H., Beeferman, D., Roy, B., and Roy, D. (2022). CommunityLM: Probing Partisan Worldviews from Language Models. In *Proceedings of the 29th International Conference on Computational Linguistics*.

Kim, J. and Lee, B. (2023). AI-Augmented Surveys: Leveraging Large Language Models for Opinion Prediction in Nationally Representative Surveys. *arXiv preprint arxiv:2305.09620*.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Lee, S., Peng, T.-Q., Goldberg, M. H., Rosenthal, S. A., Kotcher, J. E., Maibach, E. W., and Leiserowitz, A. (2023). Can large language models capture public opinion about global warming? an empirical assessment of algorithmic fidelity and bias. *arXiv preprint arXiv:2311.00217*.

Li, T., Khashabi, D., Khot, T., Sabharwal, A., and Srikumar, V. (2020). Unqovering stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics*, pages 3475–3489.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., Wang, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N., Khattab, O., Henderson, P., Huang, Q., Chi, R., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. (2022). Holistic evaluation of language models. *arXiv preprint arxiv:2211.09110*.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098.

Mao, A., Raman, N., Shu, M., Li, E., Yang, F., and Boyd-Graber, J. (2021). Eliciting bias in question answering models through ambiguity. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 92–99.

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., Miori, M., and Schmedeman, P. (2022). Do ais know what the most important issue is? using language models to code open-text social survey responses at scale. *SSRN Electronic Journal*.

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. (2018). Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391.

MosaicML (2023). Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs.

Motoki, F., Pinho Neto, V., and Rodrigues, V. (2023). More human than human: Measuring chatgpt political bias. *Available at SSRN 4372349*.

Nelson, L. K., Burk, D., Knudsen, M., and McCall, L. (2021). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, 50(1):202–237.

OpenAI (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Clark, J., Bowman, S. R., Askell, A., Grosse, R., Hernandez, D., Ganguli, D., Hubinger, E., Schiefer, N., and Kaplan, J. (2022). Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Robinson, J. and Wingate, D. (2023). Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*.

Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., and Pauly, M. (2023). The Self-Perception and Political Biases of ChatGPT. *arXiv preprint arXiv:2304.07333*.

Sanders, N. E., Ulinich, A., and Schneier, B. (2023). Demonstrations of the potential of ai-based political issue polling. *arXiv preprint arXiv:2307.04781*.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. (2023). Whose opinions do language models reflect? *International Conference on Machine Learning*.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. (2019). CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4149–4158.

Tjuatja, L., Chen, V., Wu, S. T., Talwalkar, A., and Neubig, G. (2023). Do llms exhibit human-like response biases? a case study in survey design. *arXiv preprint arXiv:2311.04076*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023a). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., and Yang, D. (2023). Can large language models transform computational social science? *arxiv preprint arxiv:2305.03514*.

A Experimental details

We use the American Community Survey (ACS) Public Use Microdata Sample (PUMS) files made available by the U.S. Census Bureau.[7] The data itself is governed by the terms of use provided by the Census Bureau.[8] We download the data directly from the U.S. Census using the Folktables Python package [Ding et al., 2021]. We download the files corresponding to the year 2019.

We downloaded the publicly available language model weights from their respective official HuggingFace repositories. We run the models in an internal cluster. The total number of GPU hours needed to complete all experiments is approximately 1500 (NVIDIA A100). The budget spent querying the OpenAI models was approximately $200.

We open source the code to replicate all experiments.[9] In addition, the repository contains notebooks to visualize the results of our investigations under different prompt ablations.

A.1 Survey questionnaire used

The exact questionnaire used in our experiments can be retrieved from our Github repository. We consider 25 questions from the 2019 ACS questionnaire corresponding to the following variables in the Public Use Microdata Sample: SEX, AGEP, HISP, RAC1P, NATIVITY, CIT, SCH, SCHL, LANX, ENG, HICOV, DEAR, DEYE, MAR, FER, GCL, MIL, WRK, ESR, JWTRNS, WKL, WKWN, WKHP, COW, PINCP. We take all questions as they appear in the ACS, with the exceptions:

- HISP: The ACS contains 5 answer choices corresponding to different Hispanic, Latino, and Spanish origins, and respondents are instructed to write down their origin if their origin is not among the choices provided. We instead provide two choices: “Yes” and “No".
- RAC1P: The ACS contains 15 answer choices, allows for selecting multiple choices, and respondents are instructed to write down their race if not among those in the multiple choice. The PUMS then provides up to 170 race codes (RAC2P and RAC3P). We instead present 9 choices, corresponding to the race codes of the RAC1P varible in the PUMS data dictionary.

Additionally, the variables ESR and COW are not directly associated with any single question in the ACS, but rather aggregate employment information. We formulate them as questions by taking the PUMS data dictionary’s variable and codes descriptions. Lastly, for the questions corresponding to the variables AGE, WKWN, WKHP, and PINCP, respondents are asked to write down an integer number. We convert such questions to multiple-choice via binning.

B Detailed experimental results

B.1 Model responses across questions before and after adjusting for A-bias

The results in this section complement Section 3, and pertain non-instruction-tuned language models. When surveying models without choice order randomization, we observe that the entropy of model responses tends to increase log-linearly with model size, often matching the entropy of the uniform distribution for the larger models. This trend is consistent across survey questions, irrespective of the question’s distribution over responses observed in the U.S. census (Figure 9).

[7] https://www.census.gov/programs-surveys/acs/microdata.html
[8] https://www.census.gov/data/developers/about/terms-of-service.html
[9] https://github.com/socialfoundations/surveying-language-models

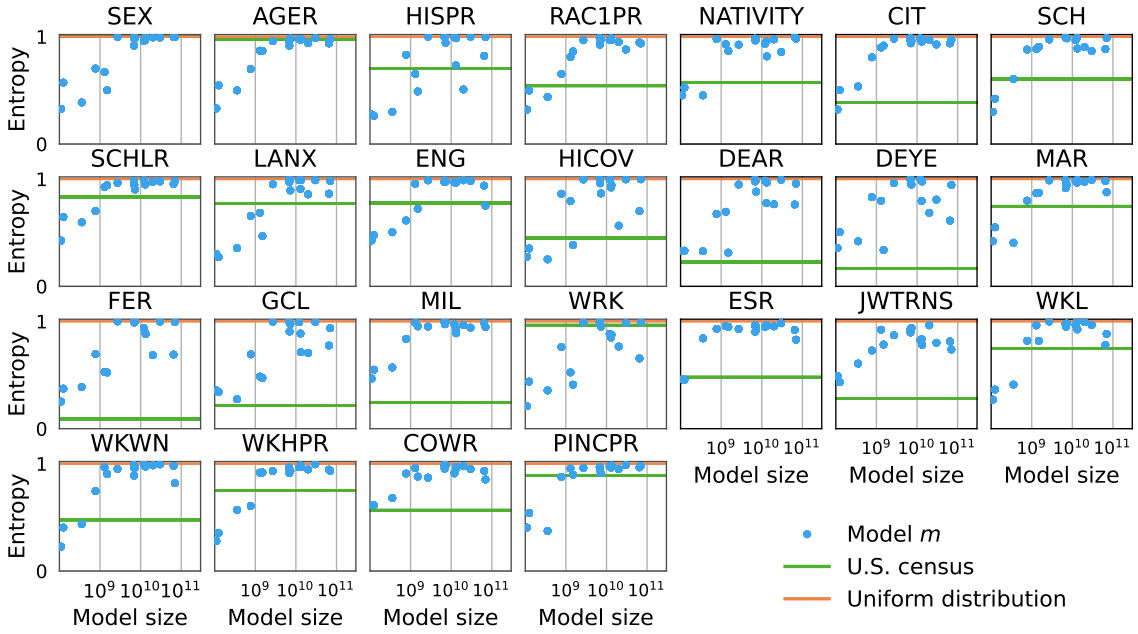


Figure 9: Normalized entropy of survey responses for individual questions (without adjustment).

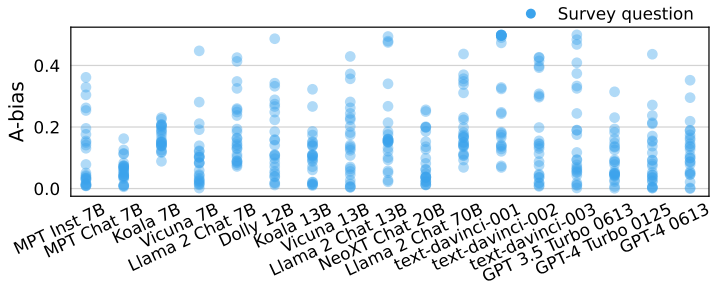


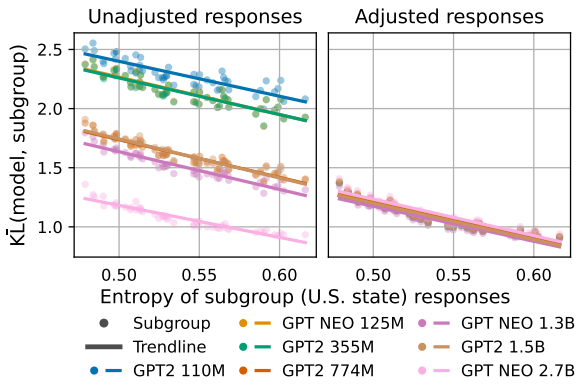
Figure 10: A-bias of instruction-tuned models.

B.2 A-bias of instruction-tuned models

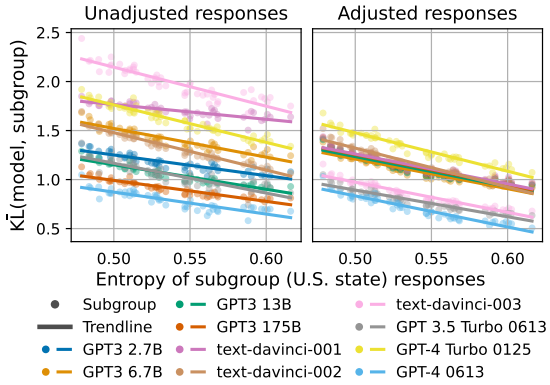
The results in this section complement Section 3.1, and pertain instruction-tuned language models as well as language models fine-tuned with reinforcement learning with human feedback (RLHF). We observe that the strength of A-bias for these models, plotted in Figure 10, is comparable to that of base pre-trained models, plotted in Figure 3(b). This motivates the use of choice-order randomization in order to eliminate confounding due to labeling biases in models' responses.

B.3 Relative alignment across demographic subgroups

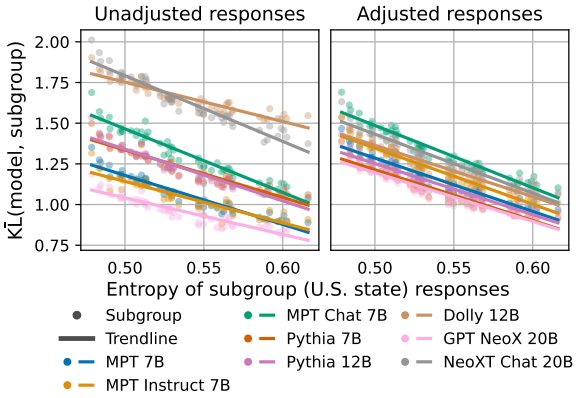
The results presented here complement those of Section 4.3. We plot the average KL divergence between each language model and each demopgrahic subpopulation (U.S. state) against the average entropy of the subgroup's responses. For readability, we split models into GPT-2 and GPT-Neo (Figure 11(a)), OpenAI's API models (Figure 11(b)), MPT, Pythia, GPT-NeoX and its instruction variants (Figure 11(c)), and LLaMA, Llama 2 and its instruction and chat variants (Figure 11(d)).



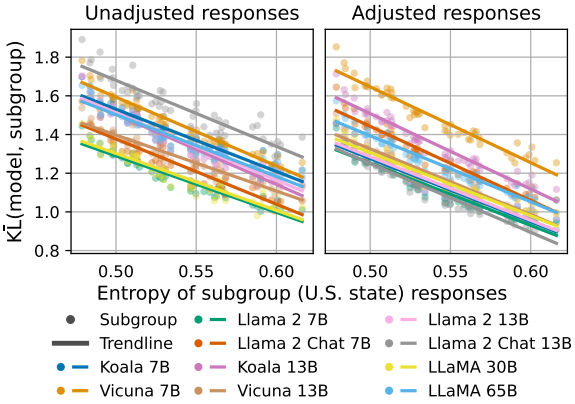
(a) GPT-2 and GPT-Neo.



(b) OpenAI's API models .



(c) MPT, Pythia, GPT-NeoX and its instruction variants.



(d) LLaMA, Llama 2 and its instruction and chat variants.

Figure 11: Relative alignment across demographic subgroups for all language models considered.

C Ordering bias: further experiments

We conduct additional randomization experiments pertaining to answer choice position and labeling bias, complimenting Section 3. We consider the GPT-2, GPT Neo, MPT, Pythia, and LLaMA models. The experiments follow a consistent setup:

1. We randomize both the order in which choices are presented and the label (i.e., letter) assigned to each answer choice. For example, for the "sex" question, the possible combinations are "A. Male B. Female", "A. Female B. Male", "B. Male A. Female", and "B. Female A. Male". Note that in the experiments presented in Section 3.1 we only randomized over the order in which choices are presented (i.e., the "A" choice was always presented first).
2. We compute the output distribution over responses for choice position (the probability assigned to the first, second, etc., answer choice presented) and letter assignment (the probability assigned to the answer choice assigned "A", "B", etc.).

For each model and survey question, we estimate the expected distribution over responses for both choice position and letter assignment by collecting 3,000 responses (step 2) under different randomizations of choice position and letter assignment (step 1). A model with no position and labeling biases would assign the same probability distribution to answer choices (e.g., "male" and "female") regardless of position or letter assignment, and therefore the expected distributions over position (e.g., selecting the first choice) and letter assignment (e.g., selecting "A") would be uniform.

C.1 Disentangling ordering bias into positioning bias and labeling bias

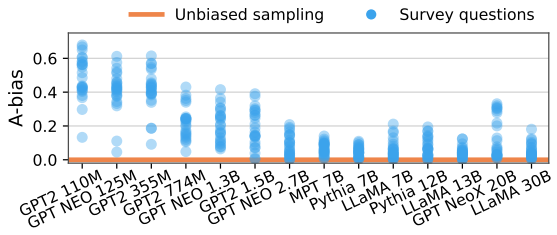
We perform chi-square tests to determine whether language models' output responses distributions over position and letter assignment significantly deviate from the uniform distribution (i.e., if there exists statistically significant bias in position or letter assignment). Since we collect 3,000 response distributions under randomized choice position and letter assignment, we ensure a high test power ($\geqslant 0.98$) in detecting small effect sizes (0.1) at a significance level of 0.05.

We find that models exhibit significant positioning and labelling for most survey questions, see Figure 12. We observe that labelling is more prevalent that positioning bias. While both tend to decrease with model size, order bias decreases more significantly with model size, whereas labeling bias tends to be very prevalent across all model sizes. In Figure 13 we plot both the strength of A-bias and first-choice bias across survey questions. The strength of A-bias tends to be greater than that of first-choice bias, particularly for the smaller models.

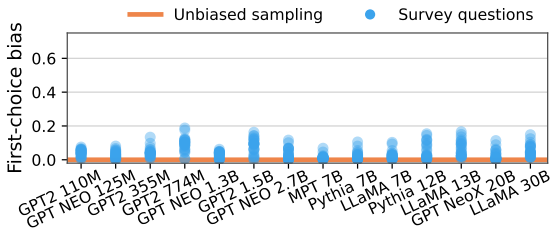
(a) Significant ordering bias.

(b) Significant label bias.

Figure 12: All models exhibit statistically significant letter and ordering bias for most survey questions.

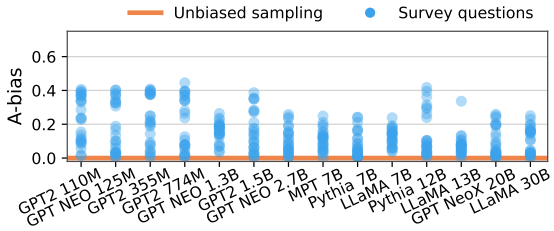


(a) A-bias

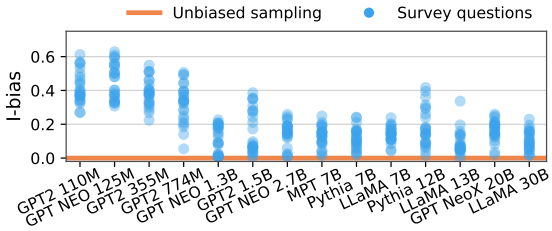


(b) First choice bias

Figure 13: Models, particularly those with less than a few billion parameters, tend to exhibit stronger A-bias than first-choice bias.



(a) A-bias in the "A", "I" randomization experiment.



(b) I-bias in the "A", "I" randomization experiment.

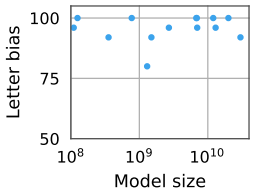
Figure 14: When both "A" and "I" are present, small models exhibit I-bias rather than A-bias.

C.2 I-bias

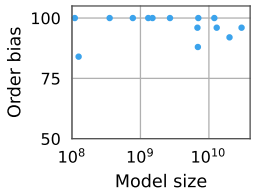
We hypothesize that A-bias is prevalent because the single character "A" is relatively frequent as the starting word of a sentence in written English. We test this hypothesis by replacing the character "B" with "I" when presenting the survey questions, since the character "I" is even more frequent as the starting word of a sentence in written English. We randomize over choice ordering and label assignment as in the previous evaluation. We find that, when presenting both "A" and "I", small models then exhibit I-bias rather than A-bias (Figure 14), supporting our initial hypothesis.

C.3 Using letters with similar frequency in written English

Motivated by the I-bias experiment, we now examine whether labeling bias can be mitigated by using letters that have similar frequency in written English. Therefore, instead of assigning to choices the labels "A", "B", etc. we assign the following labels: "R", "S", "N", "L", "O", "T", "M", "P", "W", "U", "Y", "V". We find that, compared to the "A", "B", etc. randomization experiment, the percentage of questions for which models exhibit significant labeling bias somewhat decreases (Figure 15). However, models tend to exhibit substantially more position bias. This indicates that, in the absence of a label that provides a strong signal (e.g., "A" or "I"), models tend to exhibit significantly higher choice-ordering bias, irrespective of model size.



(a) Significant labelling bias.



(b) Significant ordering bias.

Figure 15: "R", "S", "N", etc. randomization experiment. All models, irrespective of size, exhibit statistically significant letter and positioning bias for most survey questions.

D Prompt ablations

We reproduce our experiments using different prompts to query the model. Due to the cost of querying OpenAI's models, we only perform these ablations for models with publicly available weights. The notebooks with all figures can be retrieved from our Github repository.[10] Overall, the prompt ablation results are very consistent with the findings presented in the main text of the paper. In the following we provide an overview over the different ablations performed.

D.1 System rompt used for GPT-3.5 and GPT-4

When querying GPT-3.5, GPT-4, and GPT-4 Turbo, we use the system prompt `Please respond with a single letter.`, as otherwise for most questions none of the top-5 logits correspond to answer choice labels (e.g., "A", "B"). Note that this problematic arises due to the fact that the OpenAI API only allows access to the top 5 logits. We adapt the system prompt used by Dorner et al. [2023] in the context of surveying GPT-4 with standarized personality tests.

D.2 Individual survey questions

First, we use different styles to prompt individual survey questions. We enumerate the prompt styles as (P1)-(P8).

Additional context. We first explore whether including additional context signaling that the questions presented are from the American Community Survey, or that they are to be answered by U.S. households. Keeping identical survey questions, we append at the start of the prompt one of the following sentences:

(P1) `Bellow is a question from the American Community Survey.`

(P2) `Answer the following question from the American Community Survey.`

(P3) `Answer the following question as if you lived at a household in the United States.`

[10] https://github.com/socialfoundations/surveying-language-models/blob/main/prompt-ablations

Asking questions in the second person. We change the framing of the questions.

(P4) We modify the survey questionnaire such that questions are formulated in the second person rather than the third person (e.g., "What is your sex?" instead of "What is this person's sex?").

Including instructions. Following the prompt ablation of Santurkar et al. [2023], we append at the start of the prompt one of the following instructions:

(P5) `Please read the following multiple-choice question carefully and select ONE of the listed options.`

(P6) `Please read the multiple-choice question below carefully and select ONE of the listed options. Here is an example of the format:\nQuestion: Question 1\nA. Option 1\nB. Option 2\n C. Option 3\nAnswer: C`

Chat-style prompt. We consider the prompt used by Durmus et al. [2023]:

(P7) `Human: {question}\nHere are the options:\n{options}\n Assistant: If had to select one of the options, my answer would be`

Interview-style prompt. We consider the prompt used by Argyle et al. [2022]:

(P8) `Interviewer: {question}\n{options}\nMe:`

D.3 Sequential generation

We use different prompts to integrate a model's previous responses when prompting subsequent survey questions. Instead of summarizing previous responses using bullet points as in Section 5, we keep previous questions and answers in-context.

Question answering. Keeping questions and answers in-context resembles the typical few-shot Q&A setting. For instance, prompting for the third question in the questionnaire corresponds to

```
Question: {question 1}\n{options 1}\nAnswer:{answer 1}\n
Question: {question 2}\n{options 2}\nAnswer:{answer 2}\n
Question: {question 3}\n{options 3}\nAnswer:
```

Interview-style prompt. We consider the prompting style used by Argyle et al. [2022]. For instance, prompting for the third question in the questionnaire corresponds to

```
Interviewer: {question 1}\n{options 1}\nMe:{answer 1}\n
Interviewer: {question 2}\n{options 2}\nMe:{answer 2}\n
Interviewer: {question 3}\n{options 3}\nMe:
```

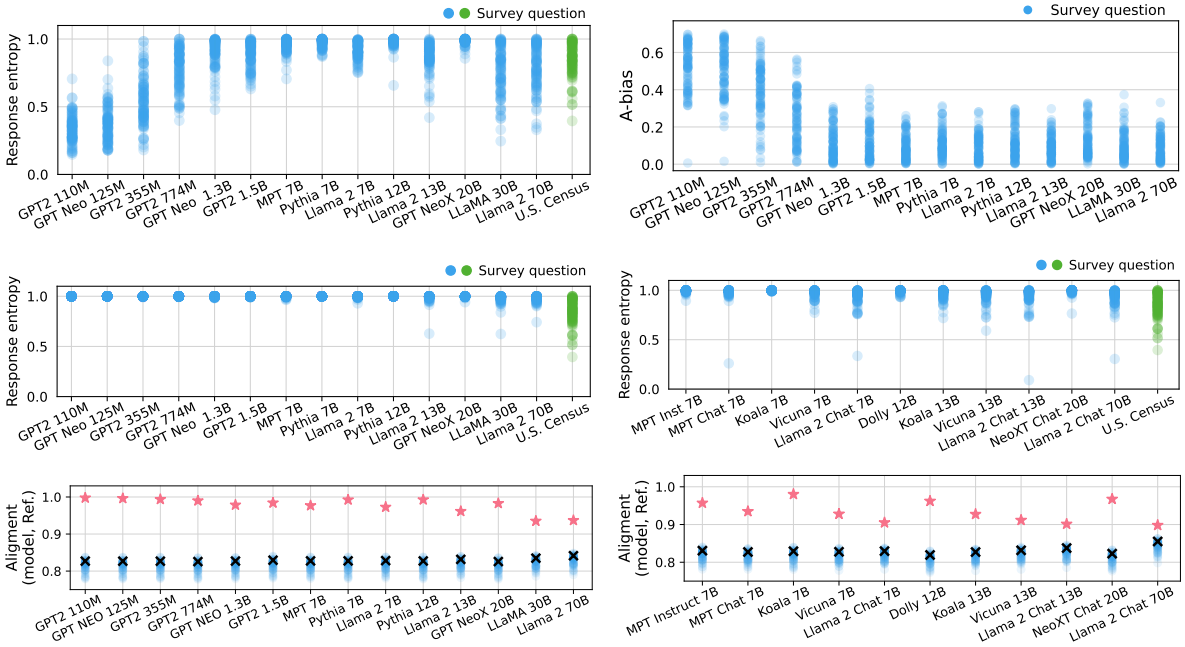


Figure 16: Reproduction of the experiments in Sections 3 and 4 for the ATP surveys.

E Results for ATP, GAS, WVS, and ANES surveys

We reproduce the experiments of Sections 3 and 4 using the ATP, and GAS/WVS used by Santurkar et al. [2023] and Durmus et al. [2023], where questions are presented individually of one another. We additionally reproduce the experiments of Section 5 using the 2016 ANES questionnaire considered by Argyle et al. [2022], where questions are presented in sequence. We do not consider OpenAI's models as the cost to reproduce the experiments via the OpenAI API exceeds our budget. We obtain very similar results to those of the ACS presented in the main text of the paper. The notebooks with all figures can be retrieved from our Github repository.[11]

E.1 ATP surveys

We obtain the ATP survey questions and their corresponding human responses from the OpinionsQA repository.[12] We present all answer choices when querying the models, but exclude the answer choices corresponding to refusals from our analysis similarly to Santurkar et al. [2023]. When comparing the similarity of models' responses to different demographic subgroups, we use the demographic subgroups and the alignment metric considered by Santurkar et al. [2023]. For such metric, higher values of alignment indicate that models' responses are more similar to the reference demographic group. We find that all models are more "aligned" with the uniformly random baseline than with any of the demographic subgroups, see Figure 16.

[11] https://github.com/socialfoundations/surveying-language-models
[12] https://github.com/tatsu-lab/opinions_qa

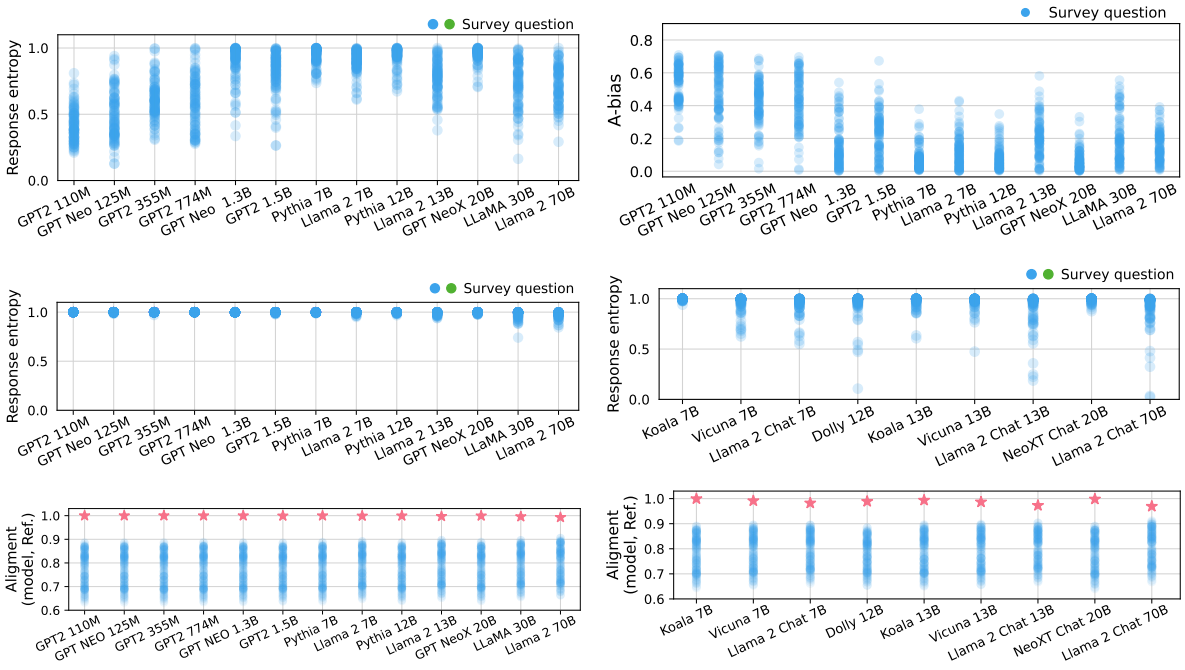


Figure 17: Reproduction of the experiments in Sections 3 and 4 for the GAS/WVS surveys.

E.2 GAS and WVS surveys

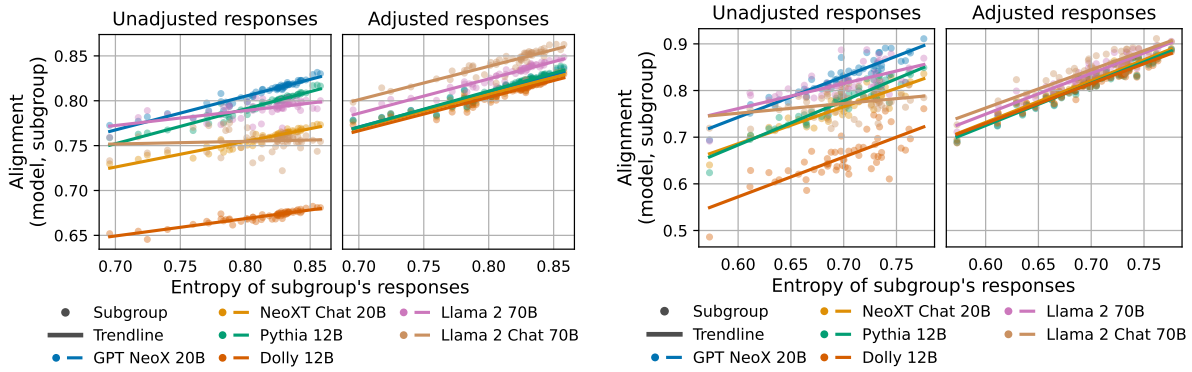
We obtain the ATP survey questions and their corresponding human responses from the GlobalOpinionsQA repository.[13] When comparing the similarity of models' responses to the population-level survey responses of different countries, we use the countries and the similarity metric considered by Durmus et al. [2023]. We find that all models produce survey responses that are more similar to those of the uniformly random baseline than to those of any of the demographic subgroups, see Figure 17.

E.3 Relative alignment for ATP and GAS/WVS surveys

We consider the alignment measures proposed by Santurkar et al. [2023] and Durmus et al. [2023] on ATP and GAS/VVS opinion surveys for the largest base / instruct models considered. We find that, similarly to our observations for the ACS, the alignment between models and a given subpopulation is highly correlated with the entropy of the subpopulations' responses.

Note that Santurkar et al. [2023] observe that RLHF can result in a "substantial shift [...] towards more liberal, educated, and wealthy [demographic groups]". Our results suggest that this could be an artifact of systematic biases. For the ATP surveys, we observe one outlier for which its alignment *before adjustment* is not correlated with the entropy of subgroup's responses: Llama 2 70B Chat, an RLHF-tuned model. However, after adjustment, Llama 2 70B Chat's alignment trend is remarkably similar to that of Llama 2 70B and all other LLMs, see Figure 19.

[13] https://huggingface.co/datasets/Anthropic/llm_global_opinions



(a) ATP surveys.

(b) GAS/WVS surveys.

Figure 18: Alignment measures proposed by Santurkar et al. [2023] and Durmus et al. [2023] on ATP and GAS/VVS opinion surveys for the largest base / instruct models considered. The alignment between models and a given subpopulation is highly correlated with the entropy of the subpopulations' responses.

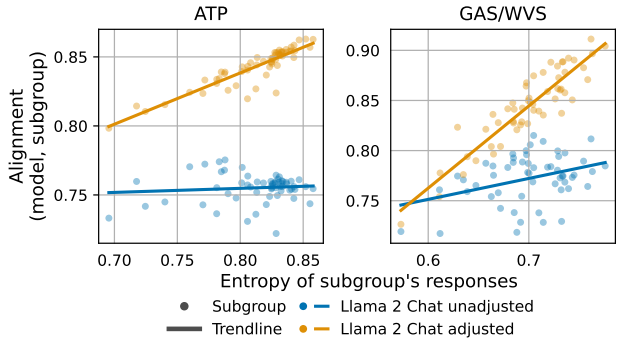


Figure 19: Alignment measures proposed by Santurkar et al. [2023] and Durmus et al. [2023] on ATP and GAS/VVS opinion surveys for Llama 2 70B Chat. The correlation between alignment and the entropy of subgroup's responses is either non-existent or weak before adjustment. However, such correlation is much stronger after adjustment, comparable to that of all other language models, see Figure 18.

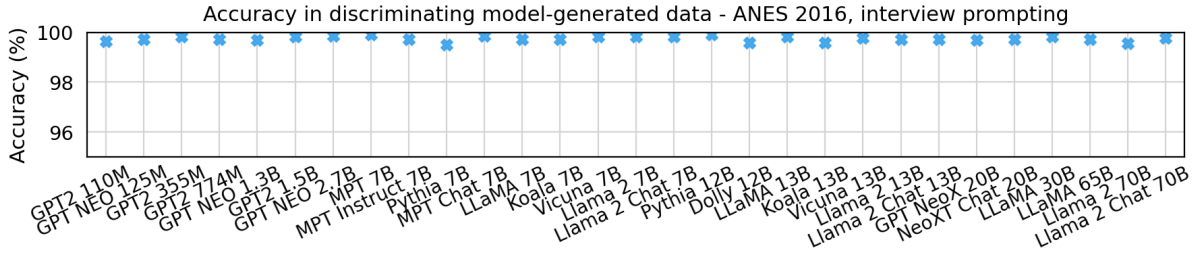


Figure 20: The discriminator test performed on datasets generated using the 2016 ANES survey questionnaire (with choice randomization).

E.4 ANES survey

We present questions in the multiple-choice format described in Section 2, using the `Interviewer:`, `Me:` prompt style described by Argyle et al. [2022]. We retrieve the 2016 ANES data from the official website[14], and process it such that it matches in form the questionnaire designed by Argyle et al. [2022]. We find that the trained classifiers can discriminate between the model-generated data and the ANES data with very high accuracy ($\geqslant$99%), see Figure 20.

[14] https://electionstudies.org/data-center/2016-time-series-study/